Baltic J. Modern Computing, Vol. 12 (2024), No. 4, pp. 646–658
https://doi.org/10.22364/bjmc.2024.12.4.24

Recent Latvian Speech Corpora for Linguistic Research and Technology Development

Ilze AUZIŅA[1], Normunds GRŪZĪTIS[1], Roberts DARĢIS[1], Guna RĀBANTE-BUŠA[1], Didzis GOŠKO[1], Jānis VEMPERS[2], Raivis KIVKUCĀNS[2], and Artūrs ZNOTIŅŠ[3]

[1] IMCS, University of Latvia, Raina blvd. 29, Riga, Latvia
[2] ZZ Dats, Ltd., Elizabetes street 41/43, Riga, Latvia
[3] RGP, Ltd., Bukaisu street 6, Riga, Latvia

ilze.auzina@lumii.lv, normunds.gruzitis@lumii.lv

ORCID 0000-0001-6143-2841, ORCID 0000-0003-0511-1829, ORCID 0000-0001-9375-6410, ORCID 0009-0005-1542-1793, ORCID 0009-0004-5528-4044

Abstract. This paper presents newly created Latvian speech corpora aimed at advancing both linguistic research and speech technology development. Although multilingual models like XLS-R and Whisper have reduced the amount of data needed for fine-tuning speech recognition models even for less-resourced languages, diverse and curated speech corpora remain essential. We provide an overview of several recent Latvian speech corpora, emphasizing their importance for both general-purpose and domain-specific use cases and comparing their design with previously created speech datasets for Latvian. We also introduce a common platform for analysing open-access Latvian speech corpora, and discuss initial evaluation and integration of speech recognition models fine-tuned on the new datasets for practical speech transcription and post-editing applications in research and industry. Finally, we present a competitive open-source speech recognition model for Latvian.

Keywords: speech corpus, general-purpose, domain-specific, corpus linguistics, language technology, Latvian language

1 Introduction

Speech corpora play a crucial role in advancing not only the language technology development – automatic speech recognition (ASR) and text-to-speech synthesis (TTS) in particular – but also the understanding and insights of phonetics and prosody, morphology and syntax, semantics and pragmatics of a language.

As demand for speech and language technology (SLT) support grows for the rapid development of various open-source and commercial applications, as well as for modern

studies in linguistics and digital humanities (DH) in general, diverse speech corpora have become even more valuable language resources for improving and evaluating ASR and TTS models, end-user applications, complex workflows, DH research and study aids – all for various needs.

Creating large curated speech corpora is a labour intensive task. Modern approaches of ASR development require less data to achieve competitive results – via fine-tuning large pre-trained multilingual models like XLS-R (Babu et al., 2022), Whisper (Radford et al., 2023) and MMS (Pratap et al., 2024). However, the importance of diverse and curated data remains and even brings new challenges in corpus creation and model evaluation.

In this paper, we present several conceptually different Latvian speech corpora recently created for both data-driven research in DH and development of ASR models for general-purpose and domain-specific use cases. Section 2 provides a brief overview of the previously created Latvian speech corpora and how they differ from the new ones, while Section 3 introduces the recent corpora. Section 4 describes the uniform platform that allows DH students and researchers to explore and analyse open-access Latvian speech corpora, and Section 5 focuses fine-tuning, evaluating and integrating ASR models for Latvian based on the recent datasets.

2 Related Work

The creation of relatively large speech corpora for Latvian began only a decade ago. Since then, several considerable speech datasets have been created, both general-purpose and domain/task-specific, primarily for the development of ASR models, secondarily for linguistic research. For instance, a diverse general-purpose 100-hour corpus LRK2013 (Pinnis et al., 2014) was the first significant resource that quickly allowed to boost the ASR support for Latvian (Salimbajevs and Strigins, 2015; Znotins et al., 2015). This dataset contains segmented orthographic transcriptions, as well as annotations of non-standard pronunciation, physiological noise, etc. It also contains a phonetic transcription layer for a 4-hour subset. Soon after, an additional 10-hour corpus of dictation and text formatting instructions was created (Pinnis et al., 2016), followed by an automatically bootstrapped 186-hour corpus of Latvian Parliament debates (Salimbajevs and Ikauniece, 2017). Domain-specific speech corpora have also been created for Latvian, notably a 35-hour corpus LVMED for the visual imaging domain (Dargis et al., 2020), which led to the first successful prototype of a medical speech transcription and post-editing system for Latvian (Gruzitis et al., 2022).

Partially following the overall design and transcription conventions of the LRK2013 corpus (Pinnis et al., 2014), a remarkable 25-hour speech corpus with orthographic transcriptions has been created also for contemporary Latgalian (MuLaR) – the largest dialect of Latvian (Juško-Štekele and Kļavinska, 2022). It documents natural, spontaneous speech, including field research recordings, interviews, TV and radio broadcasts.

Except the Latgalian corpus which is available (on request) for academic institutions as a research dataset, the rest of the above-mentioned corpora are closed datasets (proprietary or sensitive). Some of them have been recently added to the Latvian National Corpora Collection and, thus, have been made open-access (although still not as open

data) for linguistic research – for quantitative analysis via a corpus querying platform (see Section 4). On the contrary, the recently created Latvian speech corpora, reported in Section 3, are mostly available as open data or at least with an academic license.

Another aspect to mention is that these corpora were created with a pipeline ASR approach in mind, like Kaldi (Povey et al., 2011) and, later, wav2vec (Schneider et al., 2019; Baevski et al., 2020) – separate acoustic and language models, followed by a separate post-processing model to acquire a formatted transcription. Therefore, orthographic transcriptions of these corpora do not contain sentence segmentation, punctuation, abbreviations, numbers, and other text formatting. The new corpora reported in Section 3 are created with end-to-end speech transcription models like Whisper (Radford et al., 2023) in mind, which can be challenging for conversational datasets (see Sections 3.2 and 5.1).

3 Recent Latvian Speech Corpora

In this section, we present three pairs of recently created speech datasets:

- Mozilla Common Voice corpora for Latvian and Latgalian, which are the largest open speech corpora available for the two languages (Section 3.1). However, this is read speech and relatively simple language, although significant efforts have been made to include various text styles and to provoke various intonations, covering a large number of speakers.
- Two moderate-size but quality datasets representing broadcasting content and conversational speech, with the focus on rather spontaneous speech, as well as speaker diversity (Section 3.2). Both available with an academic licence for research purposes.
- Two small datasets representing the very specific language of the health domain (Section 3.3). Both are restricted-access and mostly useful for testing purposes.

3.1 Common Voice Corpora: Latvian and Latgalian

From mid 2023 to mid 2024, the Latvian part of the multilingual Common Voice corpora collection[4] was multiplied in terms of quantity and diversity. This achievement was due to the national crowdsourcing initiative BalsuTalka.lv[5] (Dargis et al., 2024b), in which a carefully selected text corpus was read by thousands of people of different ages and nationalities, both from Latvia and from the diaspora. In late 2023, this campaign was successfully launched also for Latgalian, which was not present in Common Voice before.

The first step in creating or enlarging a Common Voice (CV) speech corpus is to submit and validate a text corpus, which consists of a set of text prompts (well-formatted sentences) to be read aloud. Before the campaign, the Latvian CV corpus had around 7,000 sentences, mostly sourced from movie subtitles. To enhance the diversity of these

[4] `https://commonvoice.mozilla.org`

[5] Approximate translation of 'balsu talka': 'voice harvesting', although the concept 'talka' rather means voluntary communal work to achieve a common goal.

text prompts and increase the potential number of speech recordings, the corpus was expanded to almost 30,000 sentences. This effort not only increased the size of the text corpus but also enriched it with a much broader range of text genres, functional styles, and vocabulary. Readability and conversational style of the text fragments were prioritized, and expressive elements like questions, exclamations, dialogues, and fragments of conversations were incorporated. Continuous data augmentation was driven by the validation of recorded sentences, focusing on topic variety (e.g., news headlines, recipes), lexical diversity (including named entities), and varied sentence structures and communicative types. The CV 18.0 release[6] now includes 293 recorded hours for Latvian, with 244 hours validated, contributed by 6,086 speakers.

To create a new CV corpus for Latgalian, first, the Mozilla CV user interface was localized, and an initial set of 5,000 Latgalian sentences was selected and submitted to the CV platform (Dargis et al., 2024b). The key selection criteria were adherence to the standard Latgalian orthography, along with phonetic, intonational (narrative, questions, exclamations), and content diversity. Text snippets from dictionaries, short dialogues, and phraseology from fiction and non-fiction were manually added. The Latgalian "Bolsu tolka" campaign was organized as an extension to the Latvian "Balsu talka" campaign, which turned out to be mutually beneficial. Currently, the CV text corpus for Latgalian includes almost 10,000 sentences, while the CV 18.0 Latgalian data release contains 27 recorded hours (by 321 speakers) and 25 validated hours.

It should be once more emphasized that all the CV datasets are available as open data for both research and commercial use.

3.2 LATE Corpora: Media and Conversational

Within the State Research Programme's project LATE[7] (2022–2024), two major Latvian speech corpora have been created: LATE-Media and LATE-Conversational, more than 100 hours of recordings in total.

The LATE-Conversational corpus[8] includes recordings and orthographic transcriptions of private conversations, interviews, and public speeches. In the orthographic transcription, sentences are segmented on the basis of syntactic and prosodic cues. Pauses in the audio signal, often identified through acoustic analysis, also serve as sentence delimiters. Non-verbal elements, unclear speech, and physiological noise are annotated in the transcriptions. In addition to a normalized orthographic transcription layer where numbers and abbreviations, for instance, are expanded into full words, this corpus also contains an experimental layer of formatted transcriptions suitable for fine-tuning end-to-end speech transcription models like Whisper. Although the task of sentence splitting and text formatting is very challenging in the case of spontaneous conversational speech, it turns out to be doable if an instructed generative language model is combined with manual post-editing to transform the normalized transcription into a formatted one.

Each audio recording in the conversational corpus is also accompanied by metadata, including speaker's gender and age group (12–15, 16–25, 26–50, 51–75, 76+), as well

[6] https://commonvoice.mozilla.org/lv/datasets
[7] https://www.digitalhumanities.lv/projects/vpp-late/
[8] https://korpuss.lv/en/id/LATE-sarunas

as information about the speech type: dialogue or monologue, spontaneous or prepared speech, etc.

The LATE-Media corpus[9] includes recordings of broadcasts from Latvian public media, both spontaneous and prepared speech. The speech data is transcribed according to the orthography of the standard Latvian, following also the punctuation and other grammar rules. If necessary, annotations in square brackets indicate deviations from standard pronunciation norms (e.g. "lasām [lasam]"; "interesanti [intresanti]"), as well as pronunciation of abbreviations and foreign words (e.g. "SIA [si ā]"; "ZZS [zē zē es]"; "Rail [reil] Baltica [boltik]"), word repetitions, truncated words (e.g. "četrdesmit [čēesnt]"), and the reading of numbers, which require contextual syntactic agreement of the word forms (e.g. "7.8 [septiņi komats astoņi] grami" – nominative; "līdz 1940. [tūkstoš deviņsimt četrdesmitajam] gadam" – dative).

The size of the LATE-conversational corpus is 35 hours (more than 300 speakers), while the size of the LATE-media corpus is 70 hours (more than 250 speakers). Both datasets are distributed via the CLARIN-LV repository: LATE-media with a CLARIN Academic licence (Auzina et al., 2024a), and LATE-conversational with a CLARIN Restricted licence (Auzina et al., 2024b). Additionally, a representative LATE test set is released for benchmarking purposes (Dargis et al., 2024a).

3.3 Health Domain Corpora: Physical Rehabilitation and Visual Imaging

To adapt a speech recognition system for a specialized domain, a domain-specific speech corpus would be preferable since the vocabulary, phrases, contextual relevance, speech patterns, and technical settings can vary significantly between specific domains and use cases. However, it can often be the case that not only speech data is not available, but even a text corpus of a considerable amount is unavailable – at least for domain-specific language modelling.

Modern multilingual speech recognition architectures and models, like *wav2vec2-xls-r-300m* and *whisper-large-v3*, demonstrate significantly improved capabilities in handling out-of-domain language compared to the previous generation of ASR systems. These models, especially in their fine-tuned versions for a particular language, Latvian in our case, are more robust in adapting to diverse linguistic contexts without requiring extensive domain-specific data. Consequently, the immediate priority in specialized domains is to develop representative domain-specific test datasets, which allow for evaluation of the existing models. Only when such evaluations reveal significant shortcomings in the application of general models to specific domains and tasks (e.g., in terms of higher character, word, and formatting error rates – CER, WER, FER) should the focus shift toward creating specialized training datasets as well to further refine performance for such use cases.

To test this hypothesis in the health domain (see Section 5.1), we have selected two different subdomains – physical rehabilitation and visual imaging – for which we have created representative test sets.

A speech corpus of the physical rehabilitation domain contains prepared speech fragments, since there was no archive available with real-life dictations of rehabilita-

[9] `https://korpuss.lv/en/id/LATE-mediji`

tion reports: speech transcription (either manual or automatic) has not been part of the physical rehabilitation workflows so far in Latvia. Currently, this test corpus contains more than five hours of recordings (by more than 40 speakers) with formatted orthographic transcripts. This corpus, however, is proprietary and cannot be distributed even with an academic license.

A test set for the visual imaging domain is a balanced 1-hour subset of the above mentioned 35-hour closed-data radiology speech corpus (see Section 2), for which we have prepared formatted transcriptions. This dataset is distributed with an academic license via the CLARIN-LV repository (Znotins et al., 2024).

4 Linguistic Research

Latvian National Corpora Collection (LNCC)[10] is a diverse collection of Latvian language corpora (Saulite et al., 2022), covering both written and spoken language, and is useful (and already widely used) for linguistic studies and research, as well as language modelling. Until recently, all the spoken language corpora included in LNCC were available to Korpuss.lv users only in the form of orthographic transcriptions, i.e., as text corpora of the spoken language.

With the release of the Latvian and Latgalian Common Voice corpora (BalsuTalka and BolsuTolka), vers. 17.0, these resources are available as full-fledged speech corpora for linguistic analysis via a NoSketchEngine[11] (Rychly, 2007) instance hosted as a part of the Korpuss.lv platform. The LATE-Conversational and LATE-Media corpora have also been recently added (see Figure 1). There is also an ongoing work to include the MuLaR corpus for Latgalian (see Section 2) in LNCC.

In addition to the orthographic transcriptions and their alignment with audio segments, all Latvian speech corpora hosted on Korpuss.lv are automatically POS-tagged and lemmatized. The Latgalian speech corpus BolsuTolka, derived from the Latgalian CV dataset, is the first manually POS-tagged and lemmatized Latgalian corpus included in LNCC.

A small subset of phonetically annotated data (4 hours) has been derived from the LATE corpora (Auzina et al., 2024c).[12] The phonetic annotation is available at two levels: (1) the dictionary or standard pronunciation of a word or segment, regardless of its actual pronunciation made by the particular speaker, and (2) the actual pronunciation of a word or segment.

All speech corpora of LNCC are included in the common federated search facility of Korpuss.lv (see Figure 3) and are available for corpus linguistic analysis via the latest NoSketchEngine interface (see Figure 4). In addition to the simple search, annotation dimensions and regular expressions can be used to constrain search queries. Since LATE-Conversational is supplemented with extra-linguistic annotations, it is possible to find all occurrences of filled and silent pauses, inhalation and exhalation, laughter.

The newly created speech corpora can be used to study various phonetic and phonological phenomena of Latvian, for example, the pitch accent or syllable tone which is

[10] LNCC and its federated search platform: `https://korpuss.lv`

[11] `https://nlp.fi.muni.cz/trac/noske`

[12] `https://korpuss.lv/id/fonLATE`

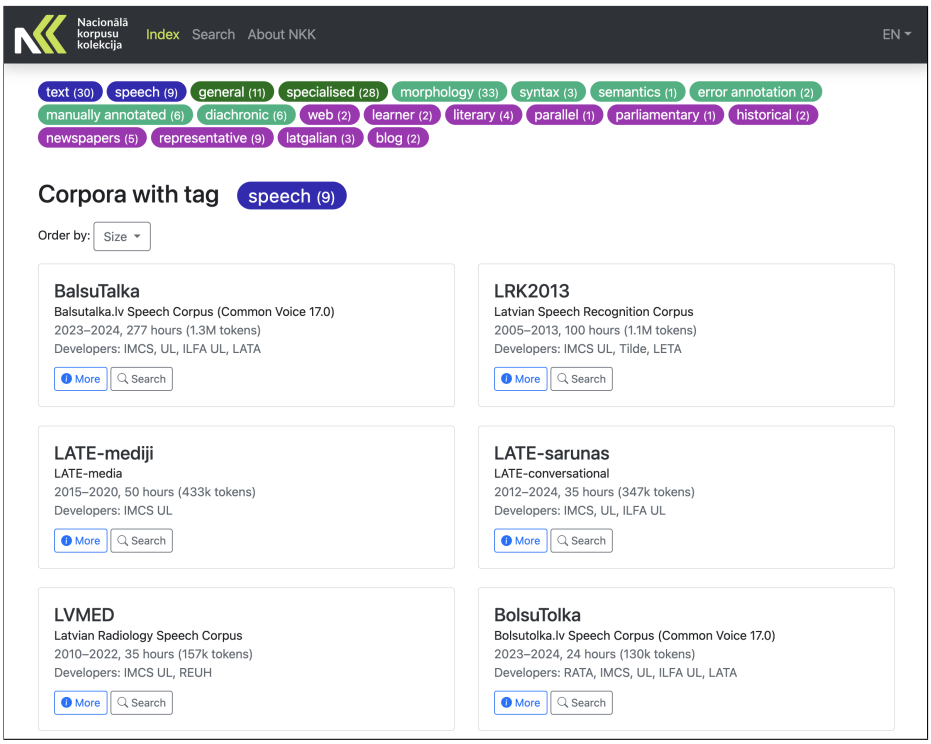

Fig. 1. Screenshot of the LNCC website (Korpuss.lv): faceted browsing of the text and speech corpora catalogue. Current selection: the largest speech corpora available for linguistic research.

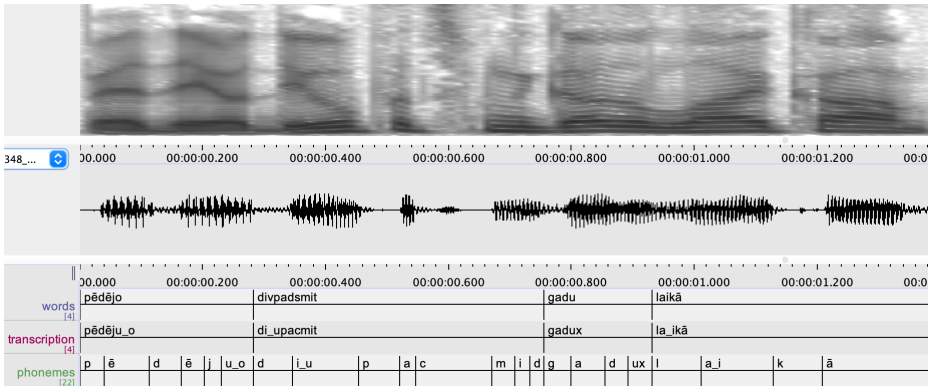

Fig. 2. Screenshot illustrating the phonetically annotated subset of LATE: the annotation layers for the phrase “pēdējo divpadsmit gadu laikā” (‘during the last twelve years’): orthographic transcription (*words*), standard pronunciation (*transcription*), and actual pronunciation (*phonemes*).

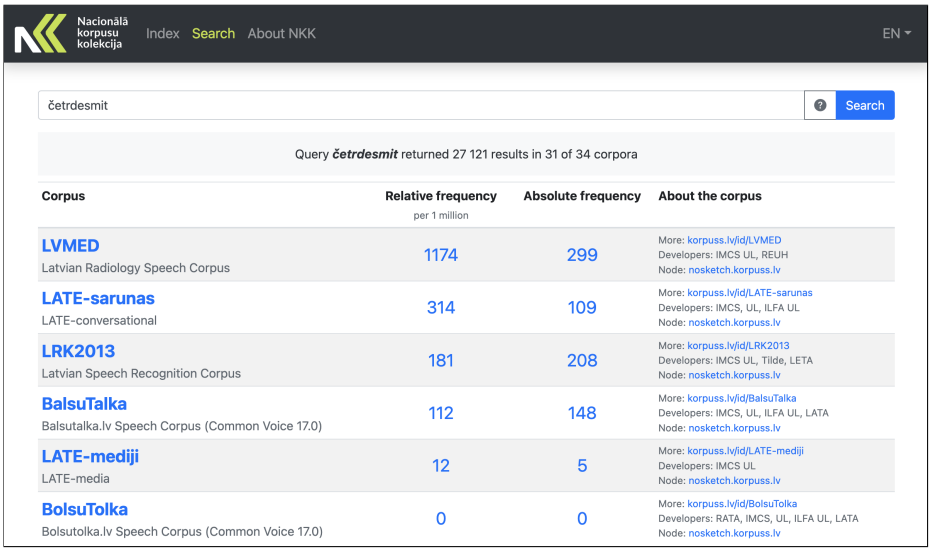

Fig. 3. Screenshot of the Korpuss.lv federated search engine: statistics of the word 'četrdesmit' ('forty', often pronounced "incorrectly" in Latvian) occurrences in LNCC speech corpora, with links to concordance lists (see Figure 4).

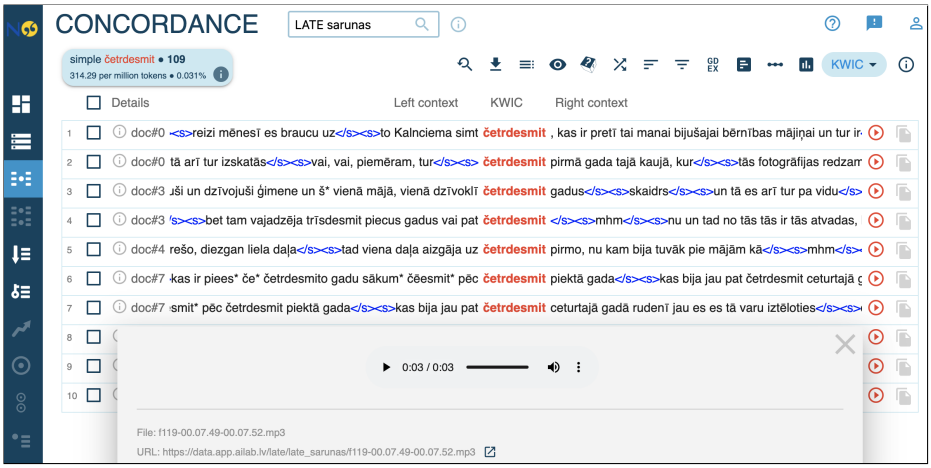

Fig. 4. Screenshot of a concordance view of the open-access NoSketch Engine corpus platform hosted at Korpuss.lv: occurrences of the word 'četrdesmit' ('forty') in the LATE-conversational corpus, aligned with the corresponding audio segments.

independent of stress and is characteristic to each long syllable. Also, the same sentences in the BalsuTalka and BolsuTolka corpora are read by several different speakers, which is valuable data to study the information structure and variation in speech w.r.t. the word order and the sentence-level intonation to mark the focus.

5 Technology Development

5.1 Open Source Models for ASR

Using the reported speech datasets (Section 3), we have fine-tuned two general-purpose multilingual ASR models – *whisper-large-v3*[13] and *mms-1b-all*[14] – for Latvian.

The fine-tuning process was conducted using Hugging Face's provided scripts for 10 epochs, with small learning rates (1e-5 for Whisper and 5e-5 for MMS), and the AdamW optimizer. A batch size of 32 was used. The datasets included the CV-19 VW split[15] and the LATE-Media training dataset. Both models were fine-tuned end-to-end without freezing any layers or using adapters.

The best fine-tuned model is available from a public Hugging Face repository[16] with the Apache 2.0 open source license.

The evaluation of the two fine-tuned models was conducted using both general-domain and domain-specific datasets. Table 1 summarizes the word error rates (WER) for each model w.r.t. each test set and compares these results to state-of-the-art baselines.

Model	**CV-19**	**LATE-Media**	**LATE-Conv.**	**Phys. Rehab.**	**Vis. Imaging**
Baseline models					
whisper-large-v3	19.2	29.1	71.6	31.0	68.9
mms-1b-all	16.0	29.3	60.6	17.7	55.9
Fine-tuned models					
whisper-large-v3-Latvian	**3.2**	**12.8**	**43.1**	**12.1**	45.9
mms-1b-all-Latvian	6.3	18.1	48.6	16.5	**45.8**

Table 1. Evaluation results (in terms of WER) of the fine-tuned and baseline models for Latvian on various test datasets (Section 3).

We used two state-of-the-art open-source models, *whisper-large-v3* and *mms-1b-all*, as baselines for our evaluation. The Whisper model performed relatively well on CV data (19.2% WER) but struggled considerably with more complex language represented by the conversational and medical corpora (WERs over 60%). The MMS model showed better generalization, significantly improving WERs on all test sets except the broadcast media dataset.

[13] `https://huggingface.co/openai/whisper-large-v3`
[14] `https://huggingface.co/facebook/mms-1b-all`
[15] `https://github.com/HarikalarKutusu/cv-tbox-split-maker`
[16] `https://huggingface.co/AiLab-IMCS-UL/whisper-large-v3-lv-late-cv19`

Fine-tuning significantly improved performance for both models, Whisper and MMS. However, the fine-tuned Whisper model significantly outperformed the fine-tuned MMS model for most test sets (see Table 1).

While fine-tuning improved general performance, results on conversational and medical domain data revealed significant challenges, with WERs still above 40%. Since WER below 10% is typically required for critical domains like legal and medical, these results highlight that further domain-specific fine-tuning is needed, particularly in the case of spontaneous speech in challenging acoustic settings or a highly specialized (domain-specific) language.

5.2 Tools for AI-in-the-Loop Speech Transcription

The fine-tuned open-source Whisper-based model has attracted attention from the industry in Latvia. It has been tested and already integrated into several AI-in-the-loop speech transcription workflows and services, e.g. to optimize the production of human-curated subtitles for social media content. It is also being further adapted for integration into municipal social services to facilitate the preparation of rehabilitation-related and other documentation.

To provide an open-source and open-access alternative, we have developed a robust and generic speech transcription tool LATE[17] (see Figure 5). It is developed with the DH community as the primary target audience in mind (including collectors of folklore and life stories, and journalists), although it is useful to the general public as well. This tool was derived from an open-source prototype previously developed for medical speech transcription (Znotins et al., 2022). First, the LATE tool segments the audio input file or stream into larger chunks using the voice activity detection method. Then these chunks are transcribed separately, which allows for parallelisation and helps to minimise hallucination issues with long audio transcription using a Whisper-based model.

A rather distinctive feature of the LATE tool is that it is available not only in the form of software as a service but also as a statically compiled and linked executable (for macOS and Linux) which together with quantized versions of the ASR models can be downloaded and run on a local PC and even on a CPU instead of a GPU.[18] Therefore, it can be used for transcribing sensitive content.

6 Conclusion

We have presented a range of recently created Latvian speech corpora, highlighting their relevance to both linguistic research and speech technology development. These datasets, which include general-purpose corpora such as Mozilla Common Voice for Latvian and Latgalian, as well as specialised corpora in health-related domains, are critical resources for evaluating and improving speech transcription systems. While modern multilingual ASR models like Whisper, MMS and XLS-R have made significant strides

[17] `https://late.ailab.lv`

[18] The GGML format and the *whisper.cpp* technology is used to achieve this: `https://github.com/ggerganov/whisper.cpp`

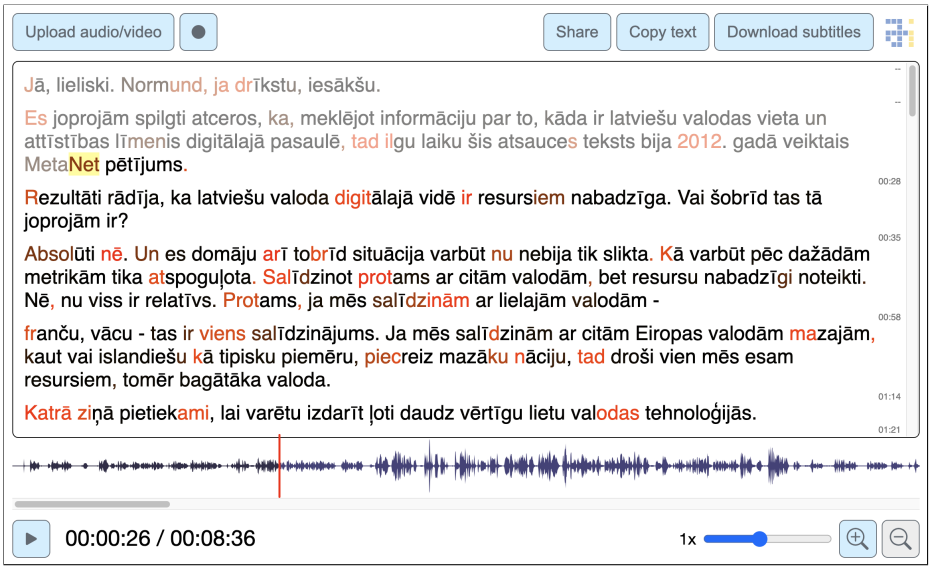

Fig. 5. Screenshot of the open-source LATE platform for automatic speech transcription and manual post-editing. It integrates the fine-tuned open-source ASR model (see Section 5.1).

in handling less-resourced languages and out-of-domain speech, domain-specific data remains crucial for further refinement and evaluation. As speech technology continues to evolve, these corpora will play a key role in ensuring that Latvian ASR systems can meet the needs of both research and practical applications across diverse linguistic contexts and domains. Based on these language resources, we have developed and released a competitive open-source ASR model for Latvian, which has been integrated in an open-source speech transcription tool for the use in digital humanities and beyond. Most of the presented speech corpora are also available (open-access) for corpus-linguistic research via the Korpuss.lv platform.

Acknowledgements

This work was funded by the State Research Programme LETONIKA, project LATE (grant agreement No. VPP-LETONIKA-2021/1-0006), in synergy with the EU Recovery and Resilience Facility, project “Competence Centre of Information and Communication Technologies” (contract No. 5.1.1.2.i.0/1/22/A/CFLA/008).

References

Auzina, I., Dargis, R., Levane-Petrova, K., Auzina, A., Saulite, B., Laksa-Timinska, I., Gailite, E., Nespore-Berzkalne, G., Rabante-Busa, G., Pokratniece, K., Klints, A. (2024a). LATE Media Speech Corpus V1 (LATE-mediji). CLARIN-LV digital library at IMCS, University

of Latvia.
http://hdl.handle.net/20.500.12574/114

Auzina, I., Dargis, R., Rabante-Busa, G., Timinska-Laksa, I., Gailite, E., Auzina, A. (2024b). LATE Conversational Speech Corpus V1 (LATE-sarunas). CLARIN-LV digital library at IMCS, University of Latvia.
http://hdl.handle.net/20.500.12574/113

Auzina, I., Rabante-Busa, G., Dargis, R. (2024c). LATE Phonetically Annotated Speech Corpus V1 (fonLATE). CLARIN-LV digital library at IMCS, University of Latvia.
http://hdl.handle.net/20.500.12574/115

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., Auli, M. (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale, *Proceedings of the 23rd INTERSPEECH Conference*, pp. 2278–2282.

Baevski, A., Zhou, Y., Mohamed, A., Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in Neural Information Processing Systems* **33**.

Dargis, R., Gruzitis, N., Auzina, I., Stepanovs, K. (2020). Creation of Language Resources for the Development of a Medical Speech Recognition System for Latvian, *Human Language Technologies - The Baltic Perspective*, Vol. 328, IOS Press, pp. 135–141.

Dargis, R., Znotins, A., Auzina, I., Rabante-Busa, G. (2024a). LATE Dev&Test Set V1 for Latvian ASR. CLARIN-LV digital library at IMCS, University of Latvia.
http://hdl.handle.net/20.500.12574/99

Dargis, R., Znotins, A., Auzina, I., Saulite, B., Reinsone, S., Dejus, R., Klavinska, A., Gruzitis, N. (2024b). BalsuTalka.lv – Boosting the Common Voice Corpus for Low-Resource Languages, *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, p. 2080–2085.

Gruzitis, N., Dargis, R., Lasmanis, V., Garkaje, G., Gosko, D. (2022). Adapting Automatic Speech Recognition to the Radiology Domain for a Less-Resourced Language: The Case of Latvian, *Intelligent Sustainable Systems*, Vol. 333, Springer, pp. 267–276.

Juško-Štekele, A., Kļavinska, A. (2022). Mūsdienu latgaliešu valodas runas korpusa izveide mazāk lietoto valodu dokumentēšanas kontekstā (Creation of Contemporary Latgalian Speech Corpus in the Context of Documenting Lesser Used Languages), *Letonica* **47**, 226–242.

Pinnis, M., Auzina, I., Goba, K. (2014). Designing the Latvian Speech Recognition Corpus, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pp. 1547–1553.

Pinnis, M., Salimbajevs, A., Auzina, I. (2016). Designing a speech corpus for the development and evaluation of dictation systems in Latvian, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pp. 775–780.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. (2011). The Kaldi Speech Recognition Toolkit, *IEEE Workshop on Automatic Speech Recognition and Understanding*.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., Auli, M. (2024). Scaling Speech Technology to 1,000+ Languages, *Journal of Machine Learning Research* **25**(97), 1–52.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision, *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Rychly, P. (2007). Manatee/Bonito-A Modular Corpus Manager, *Workshop on Recent Advances in Slavonic Natural Language Processing*, pp. 65–70.

Salimbajevs, A., Ikauniece, I. (2017). System for Speech Transcription and Post-Editing in Microsoft Word, *Proceedings of the 18th INTERSPEECH Conference*, pp. 825–826.

Salimbajevs, A., Strigins, J. (2015). Latvian Speech-to-Text Transcription Service, *Proceedings of the 16th INTERSPEECH Conference*, pp. 722–723.

Saulite, B., Dargis, R., Gruzitis, N., Auzina, I., Levane-Petrova, K., Pretkalnina, L., Rituma, L., Paikens, P., Znotins, A., Strankale, L., Pokratniece, K., Poikans, I., Barzdins, G., Skadina, I., Baklane, A., Saulespurens, V., Ziedins, J. (2022). Latvian National Corpora Collection – Korpuss.lv, *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pp. 5123–5129.

Schneider, S., Baevski, A., Collobert, R., Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition, *Proceedings of the 20th INTERSPEECH Conference*, pp. 3465–3469.

Znotins, A., Auzina, I., Saulite, B., Dargis, R., Gruzitis, N. (2024). LVMED: Test Set for Latvian ASR in the Radiology Domain. CLARIN-LV digital library at IMCS, University of Latvia. `http://hdl.handle.net/20.500.12574/117`

Znotins, A., Dargis, R., Gruzitis, N., Barzdins, G., Gosko, D. (2022). RUTA:MED – Dual Workflow Medical Speech Transcription Pipeline and Editor, *Natural Language Processing and Information Systems*, Vol. 13286, Springer, pp. 209–214.

Znotins, A., Polis, K., Dargis, R. (2015). Media Monitoring System for Latvian Radio and TV Broadcasts, *Proceedings of the 16th INTERSPEECH Conference*, pp. 732–733.

Received December 5, 2024 , accepted December 11, 2024